Human-Computer Interaction & Emerging Technologies, Vol. 195, 2025, 474–477
https://doi.org/10.54941/ahfe1006139

Novice and Expert Performance in a Knowledge Graph-Driven Assistive Dialogue System

Shannon Briggs[1], Emily Conway[2], Clare Arrington[1], Kelsey Rook[1], Tomek Strzalkowski[1], Abraham Sanders[1], Erfan El-hossami[3], and Collen Roller[4]

[1]Rensselaer Polytechnic Institute, Troy, NY, 12180 USA
[2]Air Force Research Laboratory, Wright Patterson AFB USA
[3]University of North Carolina at Charlotte, Charlotte, USA
[4]Air Force Research Laboratory, Rome, NY, USA

ABSTRACT

This paper describes a LLM-supported assistive dialog system and a proposed evaluation methodology for this system, developed for Air Force intelligence analysts in intelligence reporting scenarios. Part of this technology has previously been described in other work which developed the core technology of the assistive dialog function.

Keywords: LLM, RAG, Assistive dialogue systems, Cognitive modeling, Human-AI interaction, Expertise adaptation, Knowledge graphs

INTRODUCTION

This paper expands the previous work in merging concepts of RAG in LLMs with developed cognitive schemas represented as knowledge graphs and work flow analysis. The system used in our future user studies is a recommendation system embedded in a LLM. The dialog system is informed by knowledge graphs which describe analysts' progress through information foraging and sensemaking processes, and customizes its assistance based on that progress and expertise level of the analyst. We anticipate that this assistive dialog system will improve metrics of efficiency, effectiveness, and information absorption and retention. In order to systematize knowledge graphs for our use case, a team conducted interviews with analysts and manually recorded observations. From this aggregate of schema, we developed an ideal sample schema designed to inform the assistive dialog system to help track participants during intelligence analysis tasks.

PRIOR WORK

Existing literature is clear about expert behaviors across a variety of domains (Ericsson, 2004; Ericsson, 2013; Ericsson; 2016), and has informed our approach to developing the behavior of the assistive dialog system to support users with varying levels of expertise.

Erikson argues that long-term memory can be trained in domain-specific skills and tasks to develop long term working memory for their particular expertise, and that long term can be retrieved rapidly with specific cues. However, he suggests that long term working memory cannot be substituted for short term working memory, and is a series of states or thoughts that approximate a cognitive state. This is important to consider for our use case, as we anticipate analysts will be moving between long term and short term memory in order to navigate new information with their prior exposure to similar information and scenarios. Our goal with the dialogue agent is to help cue analysts' long term memory and allow them to access information more efficiently and ultimately perform better than without the dialogue agent.

KNOWLEDGE GRAPHS

The conversational agent's behavior during the user testing will be driven by knowledge graphs derived from domain experts. These knowledge graphs incorporate both the cognitive schema and work flow processes in order to help participants with decision making and analytical decisions.

We construct the knowledge graph based on the domain labels generated by the existing RAG system (Sanders, 2022). Domains are high-level data sources an analyst may be interested in querying. These are transformed into graph nodes and connected with weighted edges based on training dialogue data, where an edge exists if two domains have been mentioned in succession. Loops can exist, but we account for aspects like an agent asking clarifying questions and LLM suggestions by not treating these as new mentions of the domain.

We utilize the trimmed graph, which loosely serves as a cumulative dialogue state, to guide next-step suggestions in the novice mode. To achieve this, we construct a temporary subgraph consisting of the current dialogue state node, its outgoing edges, and the immediately connected nodes. A simple probabilistic rule is then applied to select the most likely consecutive node or nodes to be visited next, and pass them to the augmenting LLM to present both options to the user. Additionally, the dialogue history is leveraged to decrease the probability of revisiting nodes that were traversed in the past n dialogue steps. To tailor the baseline LLM's output for novice users, we employ Llama-3.1-8B-Instruct to modify system utterances. The level of intervention is determined by a set of heuristic rules, which dictate the appropriate prompt to use. Currently, the rules are as follows:

1. If, on the first system utterance, discussion of a specific data source has not begun (e.g. the local graph has not moved past the initial 'Start' state), the augmenting LLM is instructed to provide the user with two next-step suggestions.
2. Following the first system utterance, if a user query is resolved, we assume that progress has been made towards the high-level goal, and provide a next-step suggestion to guide the user forward.
3. In all other cases we assume that the user is either requesting general domain information or exploring a particular data source corresponding

to the current graph state. Here, direct intervention by suggesting a next step might disrupt the user's flow. Instead, we task the augmenting LLM with refining the baseline utterance to adopt a more instructional tone and include supporting information, ensuring the output is novice-friendly without interrupting the user's current task.

We use a zero-prompt prompting paradigm, but ground the augmenting LLM with brief descriptions of each data source to disambiguate out-of-distribution references.

USER STUDY

We are specifically interested in the performance difference between expert and novice subject matter experts in intelligence analysis. However, development of a large language model that adapts to user expertise has not been previously attempted. Our proposed evaluations are exploratory in nature to establish baseline behaviors.

In our evaluations, we are looking to determine if an assistive dialog system can successfully reduce training time needed to narrow the performance gap between novices and experts and if the system will have any benefit to experts in accelerating their current workflow processes. In order to study this, we have designed the dialog system's performance to differ based on the background of the participant. The dialog system will interact with experts in a support capacity, working to lessen the cognitive load of analysis. However, for novice users, the assistive dialog system engages in guiding capacity, assisting novices in tasks they have less familiarity and confidence with in order to increase efficiency, accuracy, and information retention and absorption.

Two testing scenarios have been developed to better understand the effect of conversational dialogue agents on domain-specific tasks. The first proposed testing scenario is a lab study with a dialogue agent for anticipated future interactions with participants. The second proposed testing scenario is a lab study designed to change the behavior of the dialogue agent depending on the domain level expertise of the participant. Both scenarios will monitor participants during an intelligence task they could experience in their typical workday. Participants are given up to two hours to complete an intelligence task, which covers a predetermined time frame of the proposed scenario, to gather information and deliver an intelligence estimate at the end of the testing scenario.

Anonymous surveys will be distributed to correlate an array of participant data. Pre surveys will collect standard demographic information, domain experience, as well as prior experience and comfort with automated agents. Post surveys will be distributed to capture the user's impression of the system's usefulness, utility, and efficiency. Further, based on answers to specific questions, participants will also engage in short interviews.

Success of the testing scenario is judged from a combination of human and machine metrics; human metrics being observed and collected include efficiency tasks, such as time-to-task. Other evaluation metrics include evaluation of the cognitive schema direction of the assistive dialog system, to determine if the participant is successfully directed or influenced during the course of the evaluation.

Implications and Future Research

A dialogue agent with the ability to distinguish the level of the users' domain expertise, and modifies its behavior accordingly, is a new area for LLM research. This user study examines the benefits of this assistive behavior for benefits in users' information processing and cognitive load. Based on the results from this study, we will be able to further explore how guided behavior in assisted dialogue systems can help users with differing levels of expertise. We anticipate the development of dialog agents that help decrease expert users' extrinsic cognitive load will allow this group of users to redirect task effort from database and interface management to information and analysis management. For novice users, we anticipate using the assistive dialog system to help guide users to best practices will help reduce time to task, increase information absorption, and increase overall efficiency.

CONCLUSION

This paper presents planned user studies in a novel assisted dialogue system, which merges the concept of RAGs with LLMs. This system is designed to aid intelligence analysts with standard intelligence gathering tasks, and its effect in absorption, accuracy, and efficiency will be studied among novice and expert analysts. We anticipate that the system will aid expert analysts to increase efficiency, and assist novice analysts with accuracy and information absorption.

REFERENCES

A. Sanders, "Towards a Progression-Aware Autonomous Dialogue Agent," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, preprint 2022.

Booher, Harold, ed. (2003). Handbook of human systems integration. New Jersey: Wiley.

K. A. Ericsson, *Training history, deliberate practice and elite sports performance: An analysis in response to Tucker and Collins review—what makes champions?* BMJ Publishing Group Ltd and British Association of Sport and Exercise Medicine, 2013.

K. A. Ericsson, "Deliberate Practice and the Acquisition and Maintenance of Expert Performance in Medicine and Related Domains," *Academic Medicine*, vol. 79, no. 10, p. S70, Oct. 2004.

K. A. Ericsson, "Summing up hours of any type of practice versus identifying optimal practice activities: Commentary on Macnamara, Moreau, & Hambrick (2016)," *Perspectives on Psychological Science*, vol. 11, no. 3, pp. 351–354, 2016.